Using residues coevolution to search for protein homologs through alignment of Potts models

Hugo Talibart and François Coste
Univ Rennes, Inria, CNRS, IRISA

April 30, 2019

Abstract

Thanks to sequencing technologies, the number of available protein sequences has considerably increased in the past years, but their functional and structural annotation remains a challenge. This task is classically performed in silico by retrieving well-annotated homologs with profile Hidden Markov Models (pHMMs), which are probabilistic models of families of homologous proteins capturing position-specific information with admissible insertion and deletion states. Two well-known software packages using pHMMs are widely used today: HMMER [1] aligns sequences to pHMMs to perform similarity searches, and HH-suite [2] takes it further by aligning pHMMs to pHMMs, enabling more sensitive remote homology searches.

Despite their solid performance, pHMMs are innerly limited by their positional nature. Yet, it is well-known that residues that are distant in the sequence can interact and co-evolve, e.g. due to their spatial proximity, resulting in correlated positions. Analyzing such correlations in a multiple sequence alignment by Direct Coupling Analysis [3], a statistical method to disentangle direct from indirect correlations, led to a breakthrough in the field of contact prediction [4]. Direct couplings are identified by inferring a Markov Random Field referred to as Potts model, and this model is of interest beyond its application in structure prediction. Indeed, its parameters can describe both positional conservation and direct couplings between residues of a protein. Such features drove us to examine Potts models for the purposes of modeling proteins and searching for their homologs.

In this talk, we focus on the use of Potts models for homology search, more specifically on our method for aligning and comparing Potts models. We present here our tool, named ComPotts, which formulates alignment of Potts models as an Integer Linear Programming problem and relies on a solver initially dedicated to pairwise protein alignment [5] to find efficiently the exact solution, and we present our first experimental results. Our ambition is to develop a package which would be equivalent to HH-suite but with Potts models rather than pHMMs, and to investigate on the added value of the direct couplings they provide.

References

[1] Sean R. Eddy. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.

[2] Martin Steinegger, Markus Meier, Milot Mirdita, Harald Voehringer, Stephan J Haunsberger, and Johannes Soeding. Hh-suite3 for fast remote homology detection and deep protein annotation. *bioRxiv*, page 560029, 2019.

[3] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[4] Bohdan Monastyrskyy, Daniel D'Andrea, Krzysztof Fidelis, Anna Tramontano, and Andriy Kryshtafovych. New encouraging developments in contact prediction: Assessment of the casp 11 results. *Proteins: Structure, Function, and Bioinformatics*, 84:131–144, 2016.

[5] Inken Wohlers. *Exact Algorithms For Pairwise Protein Structure Alignment.* PhD thesis, Vrije Universiteit, 01 2012.